Lec 7

* clustering

↳ Problem of finding a hidden structure within unlabeled data.

الفكرة هنا انك بتقسم الـ (data) لمجموعات، كل مجموعة متشابهة لبعضها بتكون (cluster).

* K-means clustering

ex
↳ height, weight and average lifespan of animals.

↳ used for: clustering numerical data.

Input: numerical

↳ Euclidian distance: distance must be ~~existed~~ defined over variable space.

output: Centroid

لكل (cluster) بيطلعلي مركز

~~↳ contents~~

use cases

→ introduction to classification (Discover classes)

→ exploratory technique
- ↳ Discover structure in data.
- ↳ summarize properties of each cluster.

Algorithm

① أول حاجة بتختار K بشكل عشوائي (٣ نقط مثلا)

② بعد كده بتشوف القيم الأقرب لكل نقطة منهم فيعملوا (cluster) مع بعض.

③ هتعمل إعادة حساب لل (centroid) عشان تحدد واحد جديد.

لـ هتحسب ال (mean) لكل النقط الموجودة جوه ال (cluster)

④ هتعيد الخطوة 2 و 3 لحد ما تلاقي ال (centroids) لا تتغير.

شرح ال (algorithm) كمور موجود في سلايدة ١٧ لـ ٢٨ في محاضرة ٧

ال (output) اللي بتطلع بيه ال (center) بتاع آخر (cluster).

لـ ال آخر فرق بتعرفه لل (dataset) في آخر (cluster).

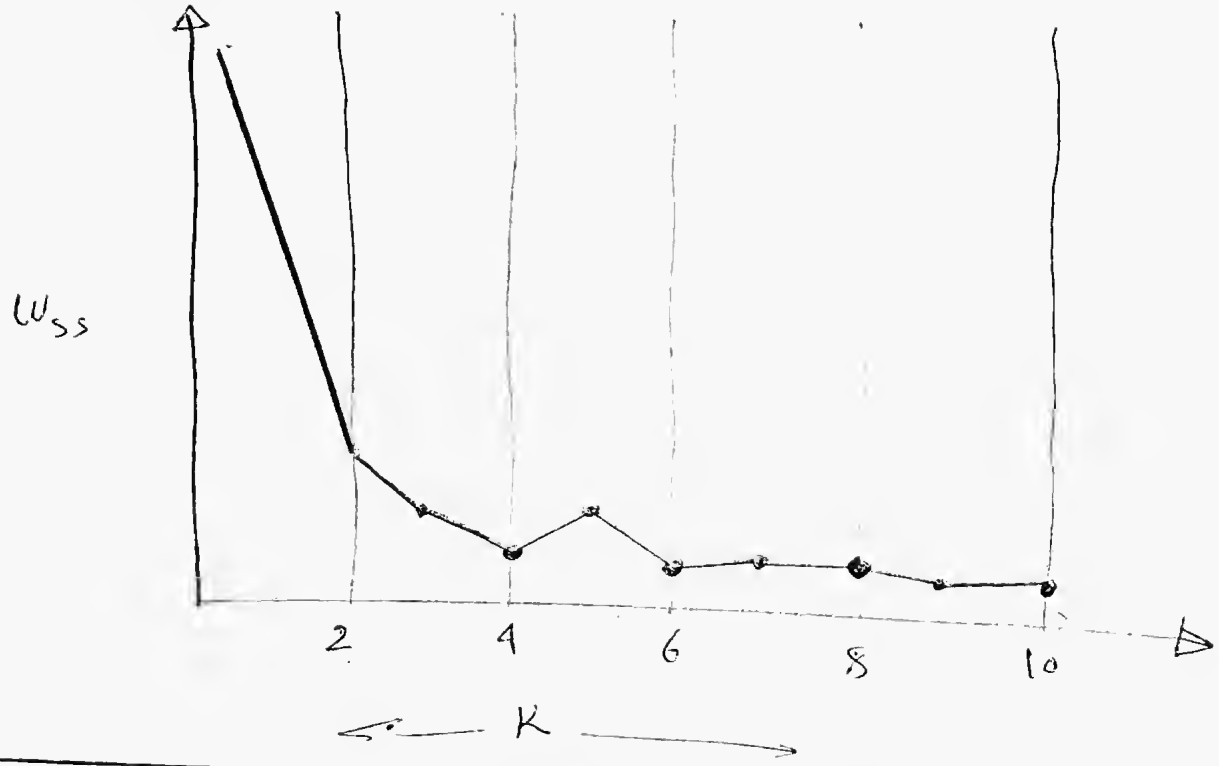
Picking K

Heuristic → Find the "elbow" of within-sum-of-squares (wss) (Plot it as function of K

$$W_{ss} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} |X_{ij} - C_i|^2$$

- K → no. of clusters
- n_i → no. of Points in ith cluster
- C_i → centroid of " "
- X_{ij} → jth Point of " "

"Elbows" at K="2, 4, 6"



- لا نيجي نقول (evaluate) لـ (model) حتكون:

1) هل الـ (clusters) باين انها متفصلة عالأقل في بعض الرسومات ولا لأ ؟

2) هل عندك (clusters) فيها (Points) قليلة لما جربت تقلل قيمة K .

3) هل فيه (centroids) قريبة من بعض لما جربت تقلل قيمة K .

K-means clustering

Reasons to choose	Cautions.
Easy to implement	Doesn't handle categorical variables.
• Easy to assign new data to existing clusters. ↳ which is the nearst cluster center	↳ sensitive to initialization (first guess)
⇝ Concise output ↳ coordinates the K cluster centers.	• not scale invariant ↳ variables should all be measured on similiar or compatible scales
	↳ Not always desirable ↳ tends to produce "round" equi-sized clusters.

Lec:8

Associan rules

↳ unsupervised learning method
↳ used to discover relationships within data.

→ ال (Association rules) مجموعة عشان تدرس على اى تغيير بيحصل فى ال (databases)

→ Apriori → works on frequent itemset (set of items that apears together).

* Apriori Property
↳ Any subset of a frequent itemset is also frequent.

مثال لو عندى itemset اسمها L عبارة عن "shoes, Purses" وهنقول ان ال (Support) = ٥٠%

لو ٥٠% من ال (transactions) ليها ال (itemset) دى
يبقى ممكن نقول ان L → frequent itemset

→ if 50% of itemsets have {shoes, Purses} in tem then at least 50% of transactions with have either {shoes} or {Purses} in them → This is Apriori Property

Lift & Leverage

Lift

بيحسب عدد المرات اللى X ، Y بيظهروا فيها معاً أكثر من المتوقع لو هما غير معتمدين على بعض.

يعنى ده قياس لكيفية X ، Y ليها علاقة ببعضهم عشان كده بيظهروا مع بعض.

$$Lift(X \rightarrow Y) = \frac{support(X \wedge Y)}{support(X) * support(Y)}$$

Leverage

هو الفرق بين احتمالية ظهور X ، Y معاً و احتماليه انهم X ، Y غير معتمدين على بعض (و ظهروا معاً بشكل عشوائى)

$$Leverage(X \rightarrow Y) = support(X \wedge Y) - support(X) * support(Y)$$

Associan rules implementations

1) Market basket analysis

2) recommender Systems.

3) discovering web usage Patterns.

⇐ مثال شرح لـ (associations rules) من ص ١١ لـ ١٢ في محاضرة رقم ٨ ⇐ فيه مسأله هتنزل في ملف تاني فيما بعد)

من حيث طاجات بتقول بيها (check) لـ (data)

1) does data make sense?
2) make "test-set" from hold-out data.
3) Evaluate rules by lift or Leverage.

Notes

*support → Percentage of transactions that contain L (set of items)

*Confidence → Percentage of transactions that contain X, which also contain Y.

*output of apriori algorithm → set of all rules X -> Y with minimum support & Confidence.

*Apriori

reasons to choose	Cautions
→ Easy to implement	*requires many database scans
→uses clever observation to Prune Search space (Apriori Property)	*EXponential time complexity.
→ Easy to Parallelize	→ Addressed with Lift and Leverage measures